MPAI: A Co-Processing Architecture with MPSoC & AI Accelerators for Vision Applications in Space

Vasileios Leon*, Panagiotis Minaidis*, Dimitrios Soudris*, George Lentaris†*

*National Technical University of Athens, School of Electrical and Computer Engineering, Zografou 15780, Greece
†University of West Attica, Department of Informatics and Computer Engineering, Egaleo 12243, Greece
Emails: *{vleon, pminaidis, dsoudris}@microlab.ntua.gr, †glentaris@uniwa.gr

Abstract—The emerging need for fast and power-efficient AI/ML deployment on-board spacecraft has forced the space industry to examine specialized accelerators, which have been successfully used in terrestrial applications. Towards this direction, the current work introduces a very heterogeneous co-processing architecture that is built around UltraScale+ MPSoC and its programmable DPU, as well as commercial AI/ML accelerators such as MyriadX VPU and Edge TPU. The proposed architecture, called *MPAI*, handles networks of different size/complexity and accommodates speed–accuracy–energy trade-offs by exploiting the diversity of accelerators in precision and computational power. This brief provides technical background and reports preliminary experimental results and outcomes.

Index Terms—On-Board Processing, Field-Programmable Gate Array (FPGA), Vision Processing Unit (VPU), Tensor Processing Unit (TPU), Deep-learning Processor Unit (DPU), Programmable Logic (PL), Deep Neural Network (DNN).

I. Introduction

The advent of the NewSpace era has been accompanied by the emergence of novel space applications, which are based on demanding AI/ML and DSP computations. As a result, traditional computing paradigms and architectures are stressed to meet the high-performance requirements in applications, among others, of Earth observation [1], vision-based navigation [2], and satellite communications [3]. To reach the performance goals, heterogeneous co-processing architectures are being examined, which also rely on Commercial Off-The-Shelf (COTS) devices [4] for the workload acceleration. At the same time, such architectures tend to meet the requirements for low power, enhanced adaptability to mission scenarios, and improved in-flight re-programmamability.

For many years, the space industry has been using FPGAs for on-board payload data processing [5]. The recent development of specialized hardware accelerators has brought AI to the fore for space. As a result, there is a plethora of works evaluating the suitability of AI accelerators, such as Edge TPU [6] and MyriadX VPU [7], to be on-board payload data processors (e.g., with respect to performance, power, cost, and radiation resilience). This radical shift for on-board processors is already apparent; spaceborne computing platforms with AI accelerators have been developed (e.g., Ubotica's CogniSAT-XE2 with MyriadX VPU and NASA's SpaceCube with Edge TPU), while there are active satellite missions with on-board AI accelerators (e.g., Φ-Sat-1 with Myriad2 VPU [1]).

In this modified landscape of on-board processing in space, the current work proposes *MPAI* (*MP*SoC + *AI*), which is a COTS co-processing architecture based on the UltraScale+ MPSoC FPGA and an AI/ML accelerator (e.g., MyriadX VPU or Edge TPU). This heterogeneous architecture can serve a great variety of workloads and tasks: instrument/sensor handling (MPSoC), general-purpose computing (MPSoC's ARM CPUs), acceleration of classic DSP functions (MPSoC's PL), and acceleration of AI/ML computations (MPSoC's PL-based DPU and/or VPU/TPU). The goal of this brief is to present the concept of MPAI, discuss key technical issues, and report preliminary comparative results. The evaluation focuses on demanding DNNs for computer vision applications in space. However, MPAI can be also used in terrestrial applications and for different types of AI networks (provided that the accelerators support their operations).

II. MPSoC-&-AI Co-Processing Architecture

The MPAI architecture, illustrated in Fig. 1, is based on MPSoC (CPU+PL+DPU) and an AI/ML accelerator (VPU, TPU, or other). MPSoC receives the camera input to be processed and handles the communication with the on-board computer.

The DPU is a softcore IP of AMD/Xilinx that implements a programmable engine in PL for inferencing DNNs. The design is based on a deep pipelined 8-bit (INT8) architecture, with the processing elements taking full advantage of the fine-grained building blocks (e.g., multipliers and accumulators). The on-chip memory is used for storing input activations, intermediate feature-maps, and output meta-data. Data reuse is applied to reduce external memory bandwidth requirements. An instruction scheduler fetches instructions from the off-chip memory to control the operation of the engine. The
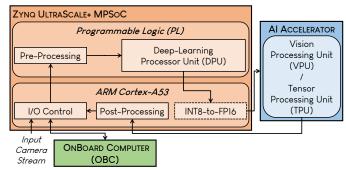

Fig. 1. Co-processing architecture with MPSoC and VPU/TPU AI accelerator.

instructions are generated by the Vitis AI compiler, which performs optimizations (e.g., layer fusion) in the network graph and generates the executable code.

The MyriadX VPU offers a great heterogeneity in processors, memories and I/O interfaces. It integrates 2 general-purpose LEON4 CPUs, 16 SIMD & VLIW programmable cores, various hardware imaging filters, and a dedicated AI accelerator engine. The memory hierarchy includes on-chip DDR DRAM, scratchpad SRAM, and caches. Specifically for AI/ML inference, the OpenVINO toolkit is used for fast deployment on the AI engine, performing model optimizations on the network's frozen graph (e.g., network pruning, linear operation fusing, and grouped convolution fusing). The models are built on 16-bit floating-point (FP16) arithmetic. MyriadX comes in two variants: SoC and USB.

The Edge TPU relies on a systolic array of multipliers & accumulators for DNN acceleration, and an on-chip SRAM for storing the model's parameters and executable. There is a USB variant that integrates only the TPU accelerator, as well as various SoM variants that add ARM processors, DDR & flash memories, and I/O interfaces. The models are first quantized to 8-bit integer (INT8) using the TensorFlow Lite toolflow, and then they are compiled with the Edge TPU compiler.

III. PRELIMINARY EXPERIMENTAL RESULTS

For preliminary evaluation, the following devices are used: (i) the ZCU104 board featuring MPSoC and implementing two instances of the DPUCZDX8G DPU, (ii) the Coral DevBoard single-board computer featuring the Edge TPU SoM, and (iii) the NCS2 USB accelerator featuring the MyriadX VPU.

The generic performance of the AI accelerators is illustrated in Fig. 2 for three networks of different complexity/size. For small networks (MobileNet V2), TPU provides $8\times$ more Frames Per Second (FPS) than VPU. However, for a larger network (ResNet-50), VPU delivers $2\times$ throughput, while for Inception V4, both accelerators sustain $\sim$10 FPS.

Table I reports results from the the acceleration of the compute-intensive UrsoNet DNN [8], which performs satellite pose estimation on the "soyuz_easy" dataset. The individual results of each accelerator include pre-processing tasks (e.g., image resampling) and DNN inference. The DPU delivers $3.8\times$ and $2.8\times$ speedup versus the VPU and the TPU, respectively. However, it is worse in terms of accuracy, as the LOCE and ORIE metrics are increased (even though TPU also uses INT8 precision). The MPAI approach (DPU +VPU) is configured using partition-aware model training: the convolutional layers are executed on the DPU with INT8, while the fully connected layers that calculate the satellite location and orientation are executed on the VPU with FP16. Namely, the demanding heavyweight layers are accelerated with the fastest DPU, while the fully-connected layers, which significantly affect the accuracy, are executed on the VPU with better precision. The MPAI latency is $2.7\times$ and $2\times$ better than that of the full execution on the VPU and TPU, respectively. Compared to the DPU, it is slightly worse, but the MPAI accuracy almost matches the baseline model accuracy.
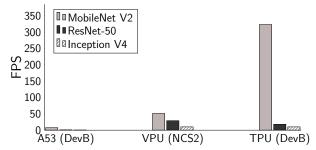

Fig. 2. Inference throughput of AI accelerators.

TABLE I
BENCHMARKING RESULTS FOR SATELLITE POSE ESTIMATION DNN
ON 1280×960×3 IMAGES

Processor / Accelerator	Hosting Device	Model Precision	Accuracy[1]		Latency	
			LOCE	ORIE	Inference	Total
Cortex-A53 CPU	DevBoard	FP32	0.68 m	7.28°	9890 ms	9928 ms
Cortex-A53 CPU	ZCU104	FP16	0.87 m	8.09°	4210 ms	4338 ms
MyriadX VPU	NCS2	FP16	0.69 m	8.71°	246 ms	252 ms
Edge TPU	DevBoard	INT8	0.66 m	7.60°	149 ms	187 ms
MPSoC DPU	ZCU104	INT8	0.96 m	9.29°	53 ms	66 ms
DPU+VPU	ZCU104 +NCS2	INT8 +FP16	0.68 m	7.32°	79 ms	92 ms

[1] Baseline SW Algorithm: · Localization Error (LOCE) = 0.63 m · Orientation Error (ORIE) = 7.20°

IV. CONCLUSION & FUTURE WORK

In this brief, an heterogeneous, mixed-precision co-processing architecture for accelerating AI/ML in space applications was presented. The advantage of this architecture is twofold: (i) it can be implemented in a single computing board integrating CPU+PL+DPU+VPU/TPU, and (ii) it efficiently accommodates various scenarios and complies with different system requirements for speed, accuracy, and energy consumption. Future work will focus on deploying virtualization & orchestration functionalities, extracting a methodology and design guidelines for the model partitioning and accelerator selection, and evaluating additional AI/ML workloads.

REFERENCES

[1] G. Giuffrida et al., "The Φ-Sat-1 Mission: the First On-Board Deep Neural Network Demonstrator for Satellite Earth Observation," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[2] V. Panousopoulos et al., "HW/SW Co-Design on Embedded SoC FPGA for Star Tracking Optimization in Space Applications," *Springer Journal of Real-Time Image Processing*, vol. 21, pp. 1–13, 2024.

[3] V. Leon et al., "Towards Enabling 5G-NTN Satellite Communications for Manned and Unmanned Rotary Wing Aircraft," in *IEEE Int'l. Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, 2024, pp. 1–6.

[4] M. C. Casey et al., "Single-Event Effects on Commercial-Off-the-Shelf Edge-Processing Artificial Intelligence ASICs," *IEEE Transactions on Nuclear Science*, vol. 70, no. 8, pp. 1716–1723, 2023.

[5] V. Leon et al., "Development and Testing on the European Space-Grade BRAVE FPGAs: Evaluation of NG-Large Using High-Performance DSP Benchmarks," *IEEE Access*, vol. 9, pp. 131 877–131 892, 2021.

[6] G. Lentaris et al., "Performance and Radiation Testing of the Coral TPU Co-processor for AI Onboard Satellites," in *European Data Handling & Data Processing Conference (EDHPC)*, 2023, pp. 1–4.

[7] V. Leon et al., "Accelerating AI and Computer Vision for Satellite Pose Estimation on the Intel Myriad X Embedded SoC," *Elsevier Microprocessors and Microsystems*, vol. 103, pp. 1–8, 2023.

[8] P. F. Proença and Y. Gao, "Deep Learning for Spacecraft Pose Estimation from Photorealistic Rendering," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 6007–6013.